Poster III-13

The Protein Information Resource (PIR)

Yeh, L.-S.*[*1], Barker, W.C.[1], Huang, H., Natale, D.A., Nikolskaya, A., Vinayaka, C.R.[1], Hu, Z.[1], Mazumder, R., Kumar, S., Suzek, B.E[1], Kourtesis, P.[1], Arminski, L.[1], Chen, Y.[1], Zhang, J.[1], Cardenas, J.[1], Chung, S., Castro-Alvear, J.[1], Wu, C.H.

Department of Biochemistry and Molecular Biology and [1]National Biomedical Research Foundation, Georgetown University Medical Center, Washington, DC, USA

The Protein Information Resource (PIR) is an integrated public bioinformatics resource that supports genomic and proteomic research and scientific studies. PIR has provided many protein databases and analysis tools to the scientific community, including the PIR-International Protein Sequence Database (PSD) of functionally annotated protein sequences. Recently, PIR has joined forces with EBI and SIB to establish UniProt (the Universal Protein Knowledgebase), the central resource of protein sequence and function, by unifying the database activities of PIR-PSD, Swiss-Prot, and TrEMBL. Central to PIR protein annotation is the PIRSF (SuperFamily) system, a network classification system based on the evolutionary relationships of whole proteins, for sensitive identification, consistent annotation, and systematic detection of annotation errors. The PIRSF database consists of preliminary clusters generated by automated classification procedures, as well as curated families with annotated information on family name, protein membership, parent-child relationship, domain architecture, and family description and bibliography. The classification supports the standardization and propagation of protein annotation, especially for position-specific functional features, protein names, keywords, and GO terms. Functional feature rules are being developed based on manually curated multiple sequence alignments and hidden Markov models of fully-curated PIRSF families with known 3D structure and experimentally verified site information. To increase the coverage of experimentally validated data, a bibliography mapping and submission system allows curators and scientists to map, categorize, and submit citations that describe the proteins.

In addition to UniProt activities, PIR distributes two databases, iProClass, an integrated database of protein family, function, and structure information, and PIR-NREF, a non-redundant reference database of protein sequences. The iProClass database provides comprehensive descriptions of all sequenced proteins and serves as a framework for data integration in a distributed networking environment. Currently consisting of about 1.1 million UniProt sequence entries organized with 36,000 PIRSF families, 6000 domains, and 1300 motifs, iProClass contains rich links and summary information for over 50 molecular databases of protein sequence, family, function and pathway, protein-protein interaction, post-translational modification, structure and structural classification, gene and genome, ontology, literature, and taxonomy. The PIR-NREF database provides a timely and comprehensive collection of protein sequence data with source attribution and minimal redundancy for sequence searching and protein identification. The database contains all sequences in PIR-PSD, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB, totaling about 1.3 million entries currently.

The PIR web site (http://pir.georgetown.edu) connects data mining and sequence analysis tools to underlying databases for information retrieval and knowledge discovery, with functionalities for interactive queries, combinations of sequence and annotation text searches, and sorting and visual exploration of search results. The FTP site provides free download for biweekly database releases in multiple formats, including XML, MySQL, and FASTA. The data integration in PIR helps users to answer complex biological questions that may typically involve querying multiple sources. In particular, interesting relationships among protein sequences, families, structures, and functions can be revealed readily. Such knowledge is fundamental to the understanding of protein evolution, structure, and function, and crucial to functional genomic and proteomic research.

The PIR is supported by grant U01 HG02712 from the National Institutes of Health, and grants DBI-0138188 and ITR-0205470 from the National Science Foundation.